Welcome to Advance Workshop
Big Data Analytic and AI/ML on AWS

Rudi Suryadi, Yudho Diponegoro
AWS Solutions Architect, Indonesia

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

What do you expect from this Workshop?

1. **Big Data Analytic**
    1. **AWS Big Data Services**
    2. **Building Serverless Data Lake**
        - Lab1: Building Data Lake on AWS (Glue, Athena, Quicksight)
    3. **Modern Data Architecture**
        - Lab2: Real Time Data Processing
2. **AI/ML**
3. **Quiz – Refresh your Learning Today**
4. **What's Next**
5. **Feedback**

20% Theory
80% Hands-on

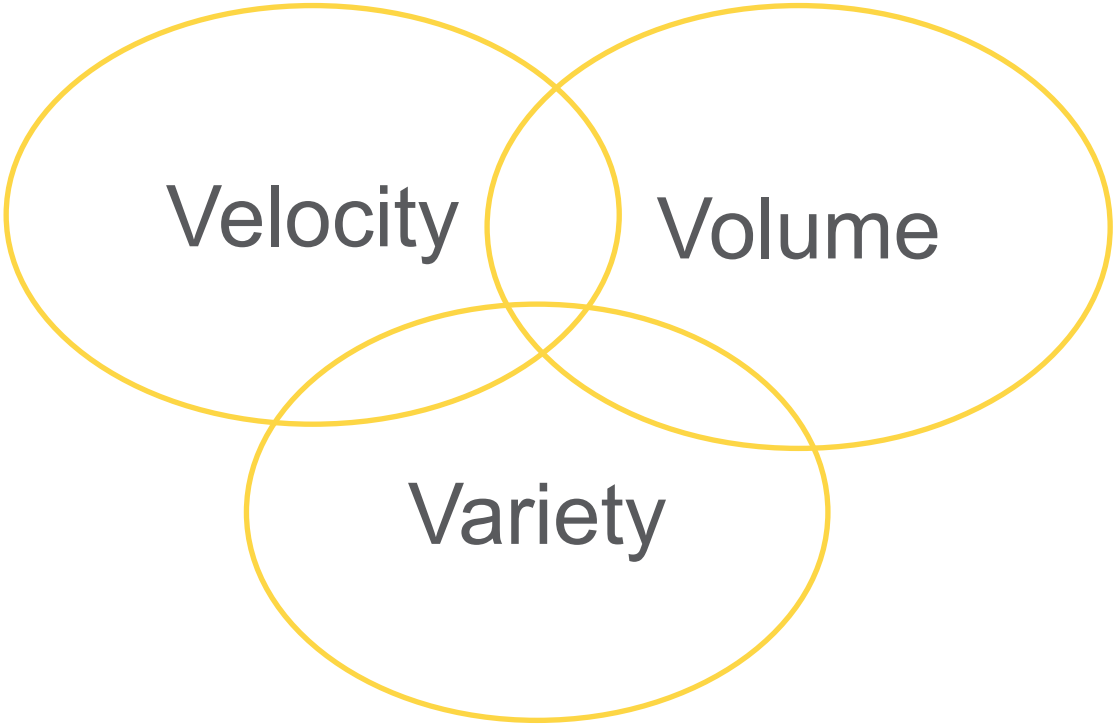
aws

Workshop Requirement:

- Laptop
- Use provided AWS account
- Basic understanding of AWS cloud
- Basic Understanding of SQL
- Basic Understanding of Python
- Willingness to Hands-on

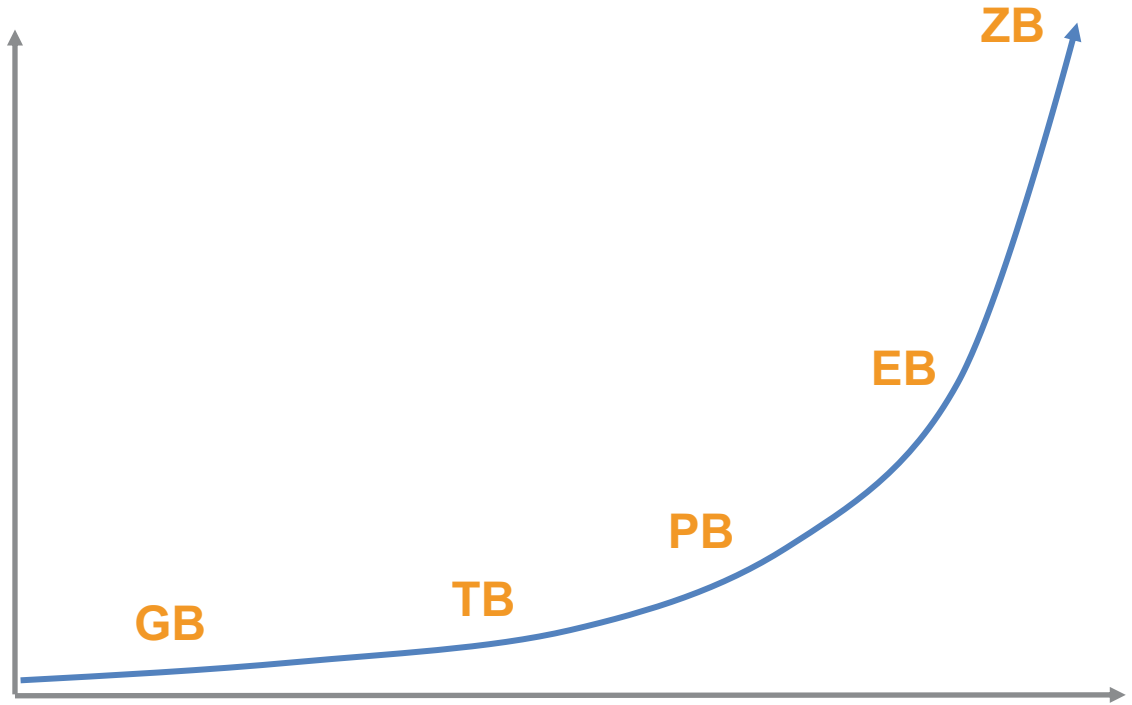
aws

Three Vs of Big Data



aws

Big Data: Unconstrained Growth



Unstructured data growth is explosive

95% of the 1.2 zettabytes of data in the digital universe is unstructured

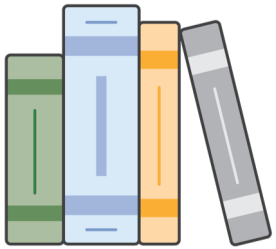
Machine data and IoT will only steepen the curve

70% of this data is user-generated content

Source: IDC, The Internet of Things: Getting Ready to Embrace Its Impact on the Digital Economy, March 2016.

aws

Big Data Sources



Sources

- Relational
- NoSQL
- Web servers
- Mobile phones/Tablets
- 3rd party feeds
- IoT
- Clickstream

aws

Big Data Formats and Velocity

Formats

Structured

Unstructured

Text

Binary

Velocity

Real-time/Near Real-time

Batched

aws

Why Big Data?

Get answers faster and be able to ask questions not possible to today.

Security threat detection

User Behavior Analysis

Smart Application (Machine Learning)

Business Intelligence

Fraud detection

Financial Modeling and Forecasting

Spending optimization

Real-time alerting

aws

Elastic and highly scalable

+

No upfront capital expense

+

Only pay for what you use

+

Available on-demand

= **the Cloud removes constraints**

aws

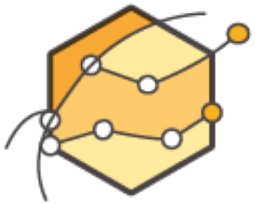
The Cloud Was Built for Big Data

Agility: Try more, fail fast, go big or start small, and process data at any scale

Scalability: Run jobs any time, without guessing capacity or limiting functionality

Broadest and Deepest Capabilities: Access 70+ managed Big Data services to address any workload

Low Cost: Pay only for the IT you use, when you use it

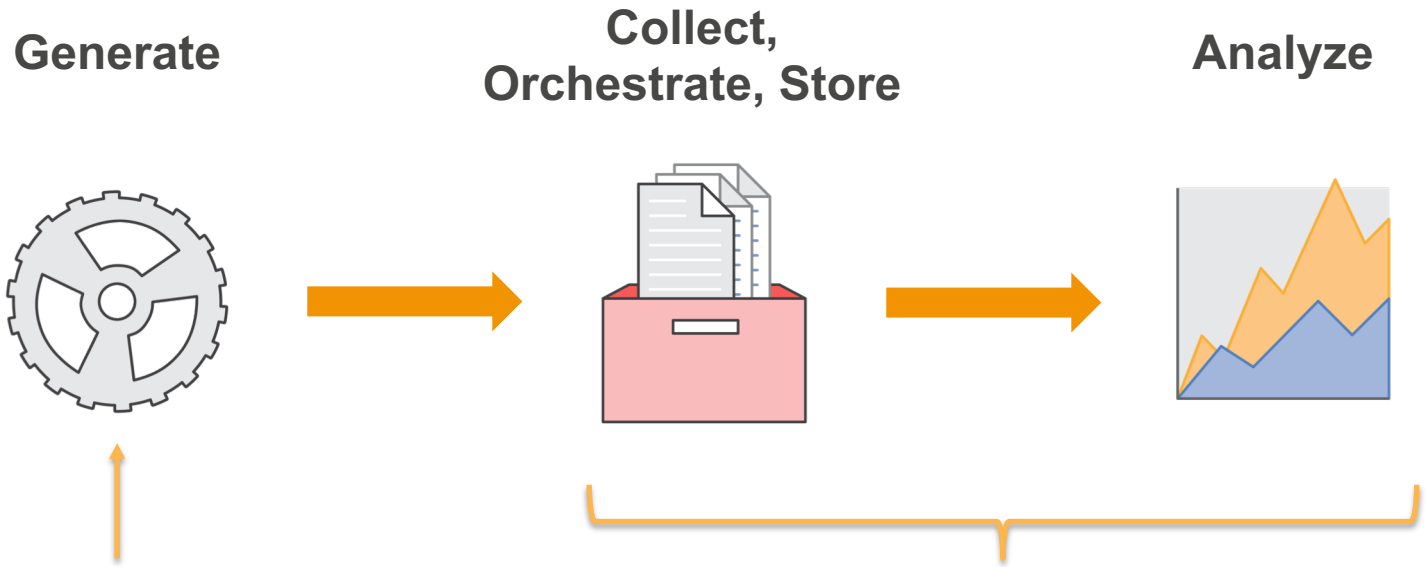


Get to Insights Faster: Focus on data science not the heavy undifferentiated lift of managing raw data

Data Migrations Made Easy: Move exabyte-scale data to the cloud quickly and cost-effectively

aws

Common Big Data Flow



AWS Big Data Services

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

AWS Big Data Platform

Collect

Direct Connect

Import Export

AWS Snowball

AWS IoT

Kinesis

AWS Database Migration Service

Orchestrate

AWS Lambda

AWS Data Pipeline

Amazon SNS

Amazon SWF

AWS Glue

Store

S3

Glacier

DynamoDB

Amazon Aurora

Analyze

EMR

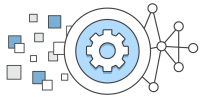
EC2

Redshift

Machine Learning

Amazon Kinesis

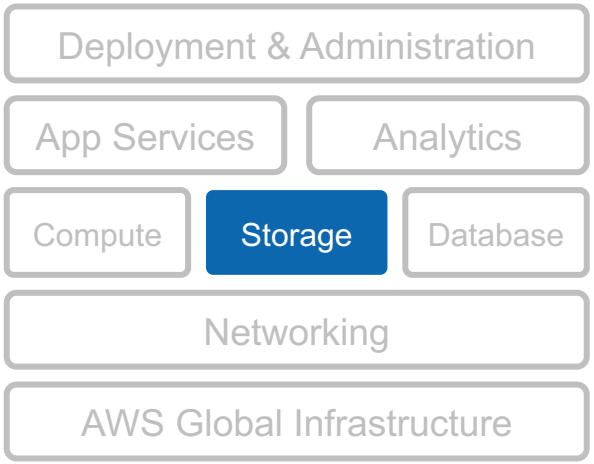
Amazon QuickSight



Amazon Athena

aws

Amazon S3



Scalable object storage for the Internet

1 byte to 5 TB in size per object + unlimited number of objects

99.999999999% durability, 99.99% availability

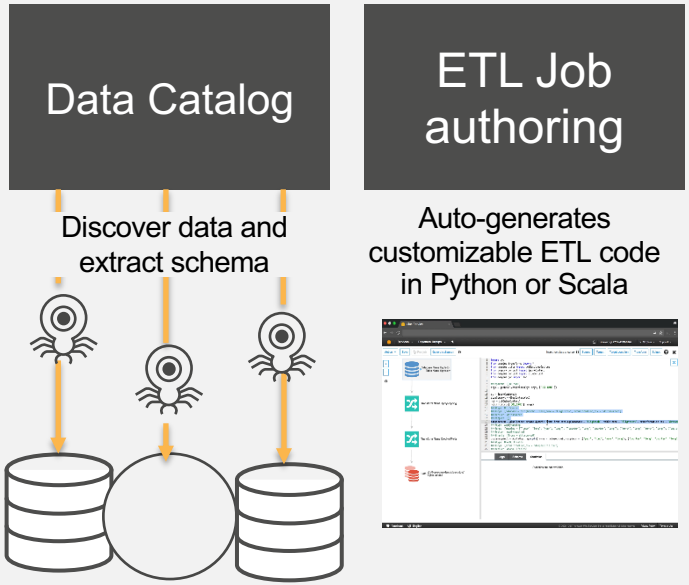
Regional service, no single points of failure

Server side encryption

aws

AWS Glue

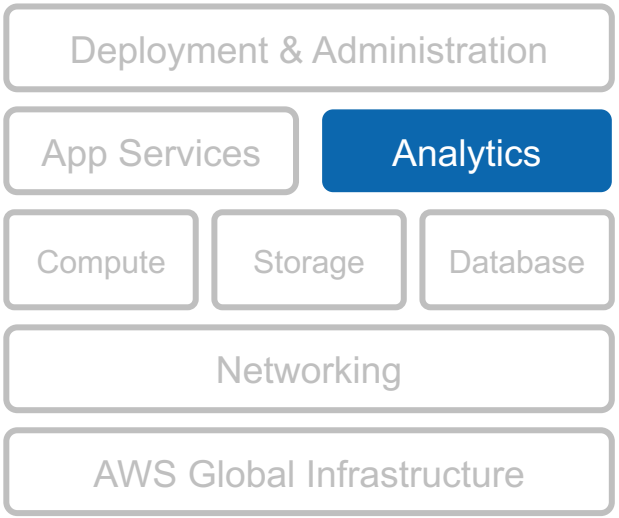
Serverless Data Catalog & ETL Service



- Automatically discovers data and stores schema
- Data is immediately searchable, and available for ETL
- Automatically generates customizable code
- Schedules and runs your ETL jobs
- Serverless

© 2018, Amazon Web Services, Inc. or Its Affiliates. All rights reserved.

Amazon Athena



Query and analyze Amazon S3 data with standard (ANSI) SQL

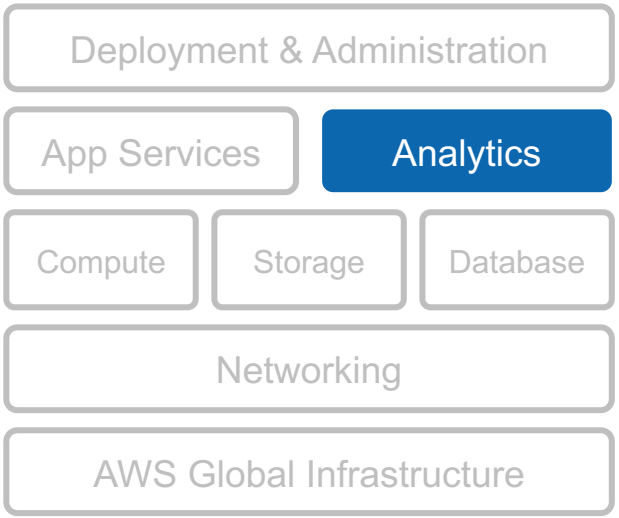
No ETL required

Serverless and simple

Pay only for queries you run

aws

Amazon QuickSight



BI service performs ad-hoc analysis

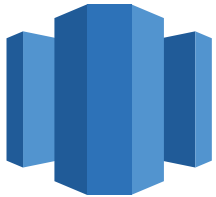
Build visualizations

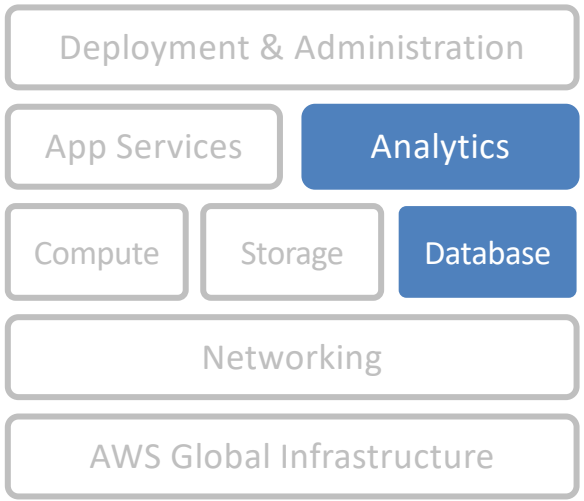
Share and collaborate via storyboards

Native access on major mobile platforms

aws

Amazon Redshift





Managed Massively Parallel Petabyte Scale Data Warehouse

Streaming Backup/Restore to S3

Load data from S3, DynamoDB and EMR

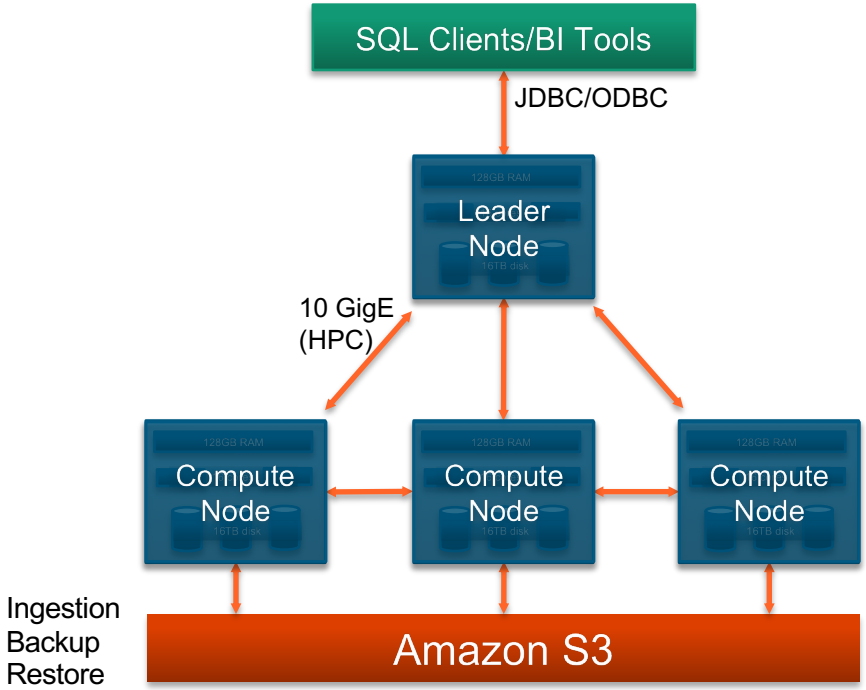
Extensive Security Features

Online Scaling from 160 GB -> 2 PB

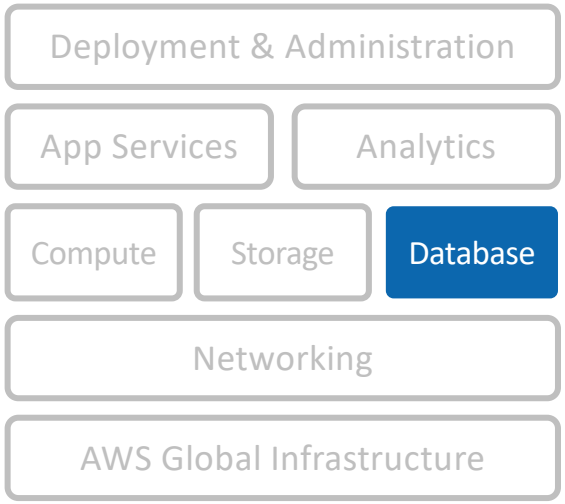
aws

Amazon Redshift

- **Scalability & Elasticity**
  - Resize or scale - Number or type of nodes can be changed with a few clicks
- **Durability and Availability**
  - Replication
  - Backup
  - Automated recovery from failed drives & nodes
- **Interfaces**
  - JDBC/ODBC interface with BI/ETL tools
  - Amazon S3 or DynamoDB
- **Anti-patterns**
  - Small datasets (smallest database 160GB)
  - OLTP
  - Unstructured Data
  - Blob Data



Amazon DynamoDB



Fully managed NoSQL database

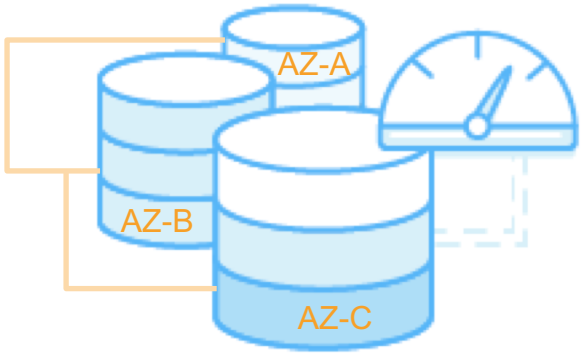
Single-Digit Millisecond latency at scale

Supports document and key-value

aws

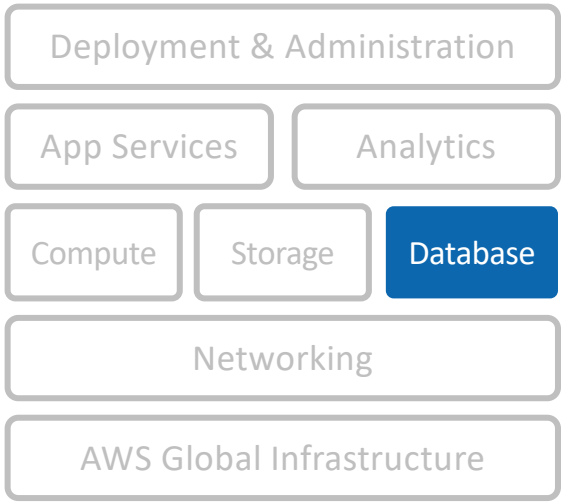
Amazon DynamoDB

- **Durability and Availability**
  - Three Availability Zones (AZ)
- **Interfaces**
  - AWS Management Console
  - API's
  - SDK's
- **Anti-patterns**
  - Application tied to traditional relational database
  - Joins and or complex transactions
  - BLOB data
  - Large data with low I/O rate



aws

Amazon Aurora



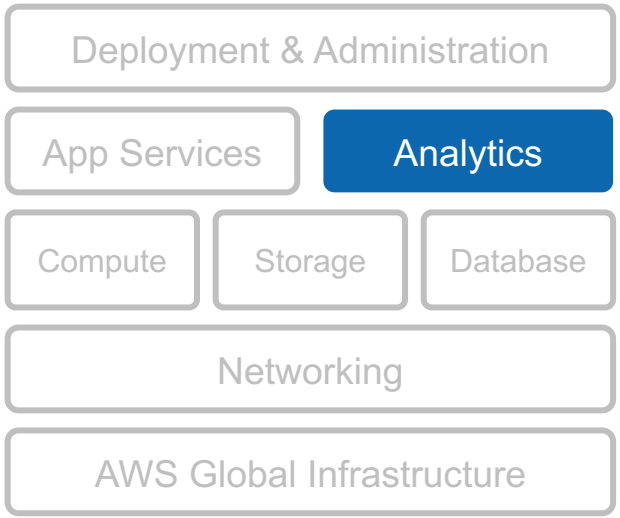
5x performance at 1/10th the cost of alternatives

Fully managed MySQL-compatible database

Fast with 500K reads/100K writes per second

aws

Amazon Kinesis



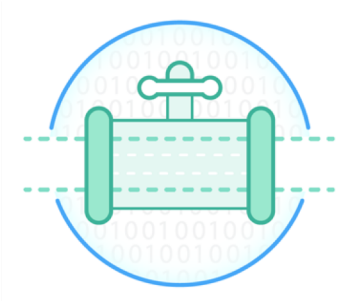
Ingest streaming data

Process data in real-time

Store terabytes of data per hour

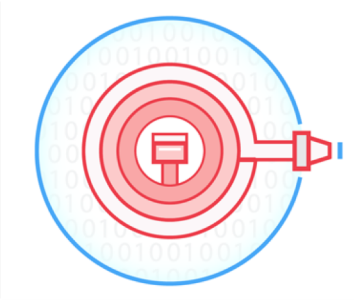
aws

Amazon Kinesis



Amazon Kinesis Streams

Build your own custom applications that process or analyze streaming data



Amazon Kinesis Firehose

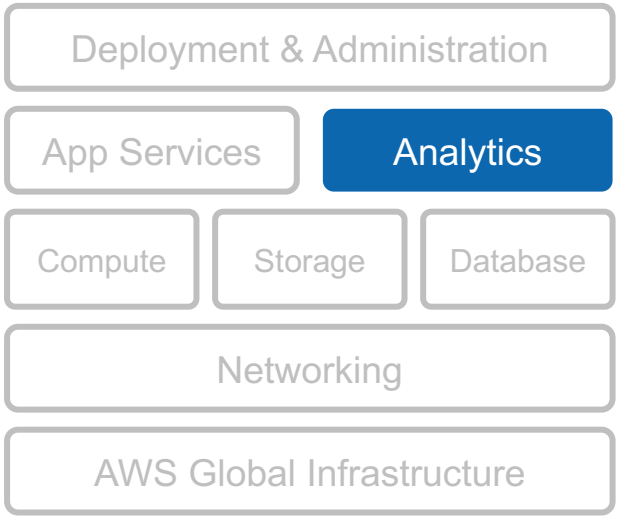
Easily load massive volumes of streaming data into Amazon S3 and Redshift

Amazon Kinesis Analytics

Easily analyze data streams using standard SQL queries

aws

Amazon EMR



Scalable Hadoop/Spark clusters as a service

Launch a cluster in minutes

Hadoop, Hive, Spark, Presto, HBase, etc.

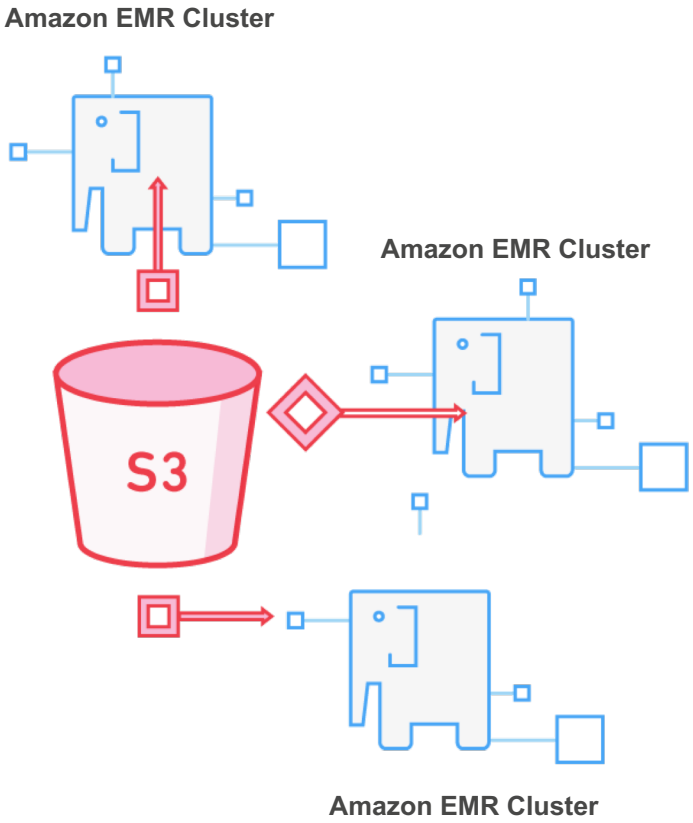
Easy to use; fully managed

HDFS, Amazon EBS, and S3 file systems

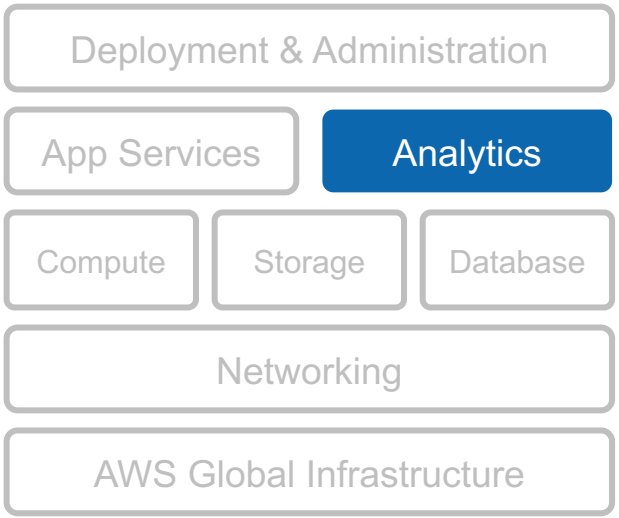
aws

Amazon EMR

- **Scalability & Elasticity**
  - Resize a running cluster based on how much work is needed to be done.
- **Durability and Availability**
  - Fault tolerant for slave node (HDFS)
  - Backup to S3 for resilience against master node failures
- **Standard Interfaces**
  - Hive, Pig, Spark, Hbase, Impala, Hunk, Presto, other popular tools



Machine and Deep Learning



Amazon Machine Learning
scalable and robust implementations of industry-standard ML supervised learning algorithms

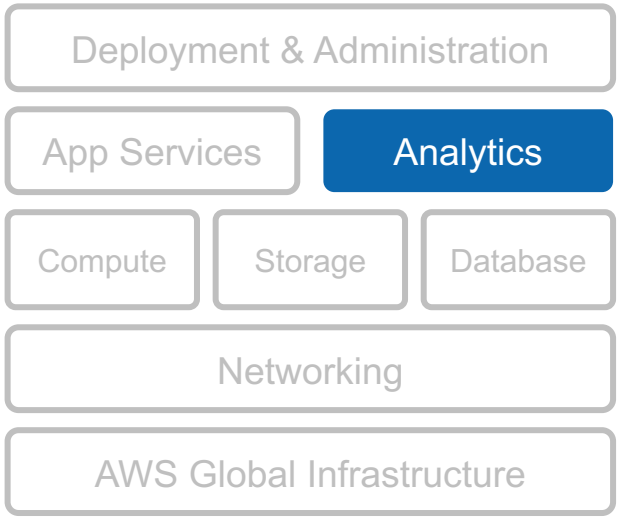
Amazon Lex
Conversational interfaces through Voice or Text
Backend powering Alexa

Amazon Polly
Cloud Native TTS (Text to Speech)
47 lifelike voices/24 languages (on growing)
Low-latency for real-time applications

Amazon Rekognition
Deep learning-based image recognition
Object/Scene detection, facial analysis and comparison

aws

Amazon Elasticsearch Service



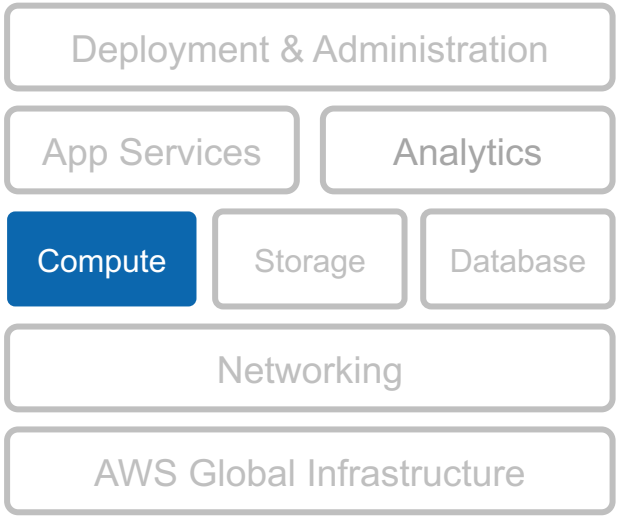
Setup Elasticsearch cluster in minutes

Integrated with Logstash and Kibana

Scale Elasticsearch clusters seamlessly

aws

Amazon EC2



Scale up or down as needed

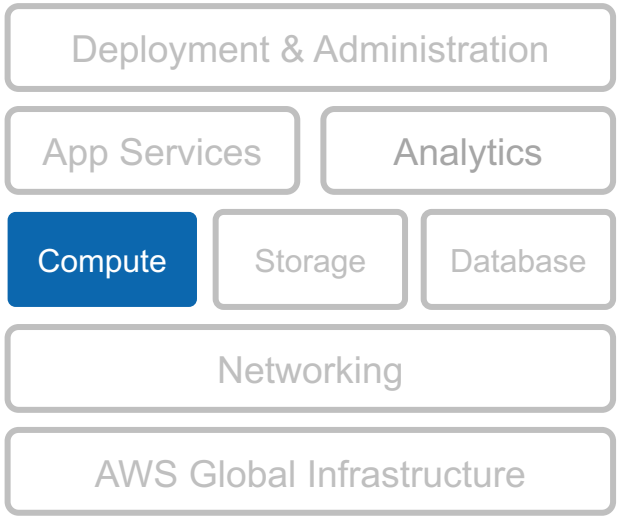
Pay for what you use

Largest select of instance types

Do-it-yourself big data applications

aws

AWS Lambda

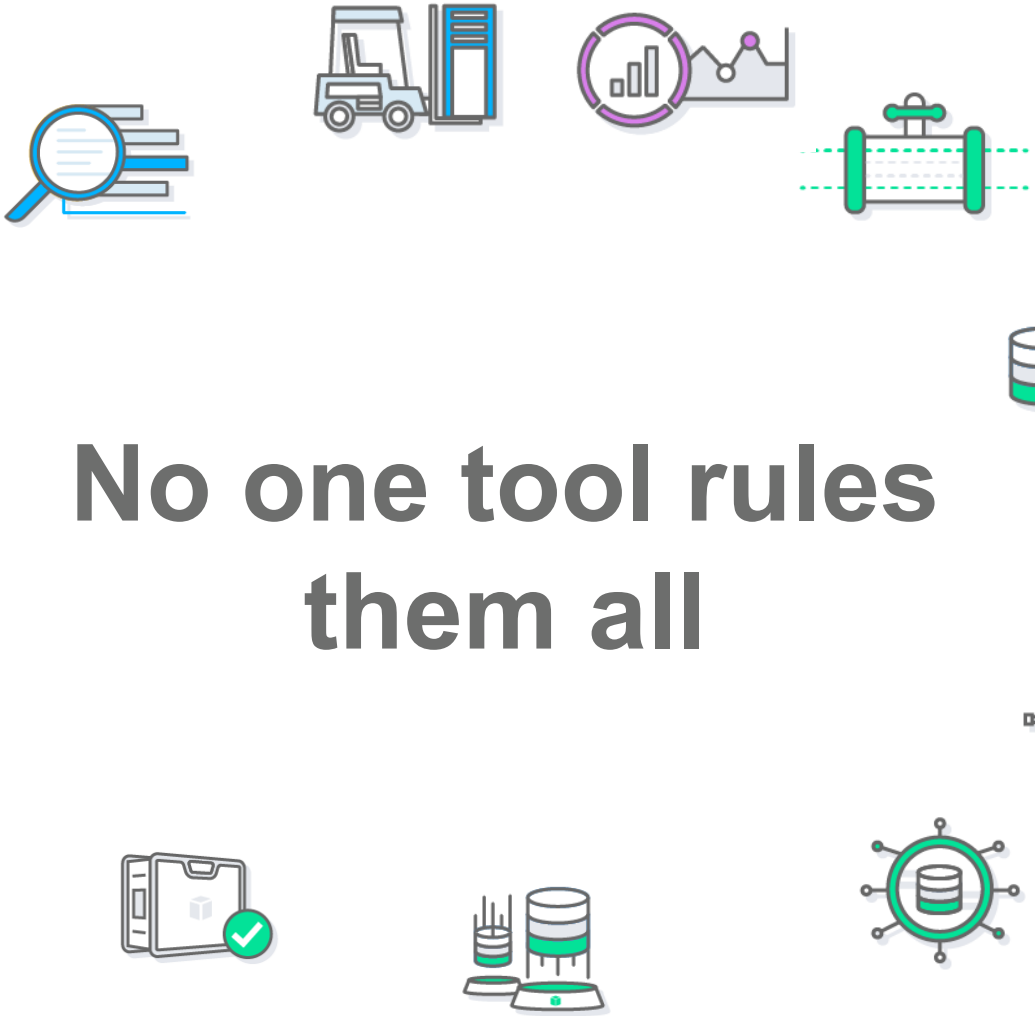


Event driven, fully managed compute

No Infrastructure to Manage

Automatic Scaling

aws


No one tool rules them all

The AWS Approach

- Flexible - Use the best tool for the job
    - Data structure, latency, throughput, access patterns
- Low Cost - Big data ≠ big cost
- Scalable – Data should be immutable (append-only)
    - Batch/speed/serving layer
- Minimize Admin Overhead - Leverage AWS managed services
    - No or very low admin
- Be Agile – Fail fast, test more, optimize Big Data at a lower cost

aws

Data Lake on AWS

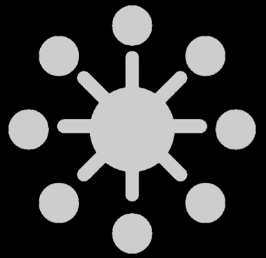
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

Characteristics of a Data Lake

Collect Anything

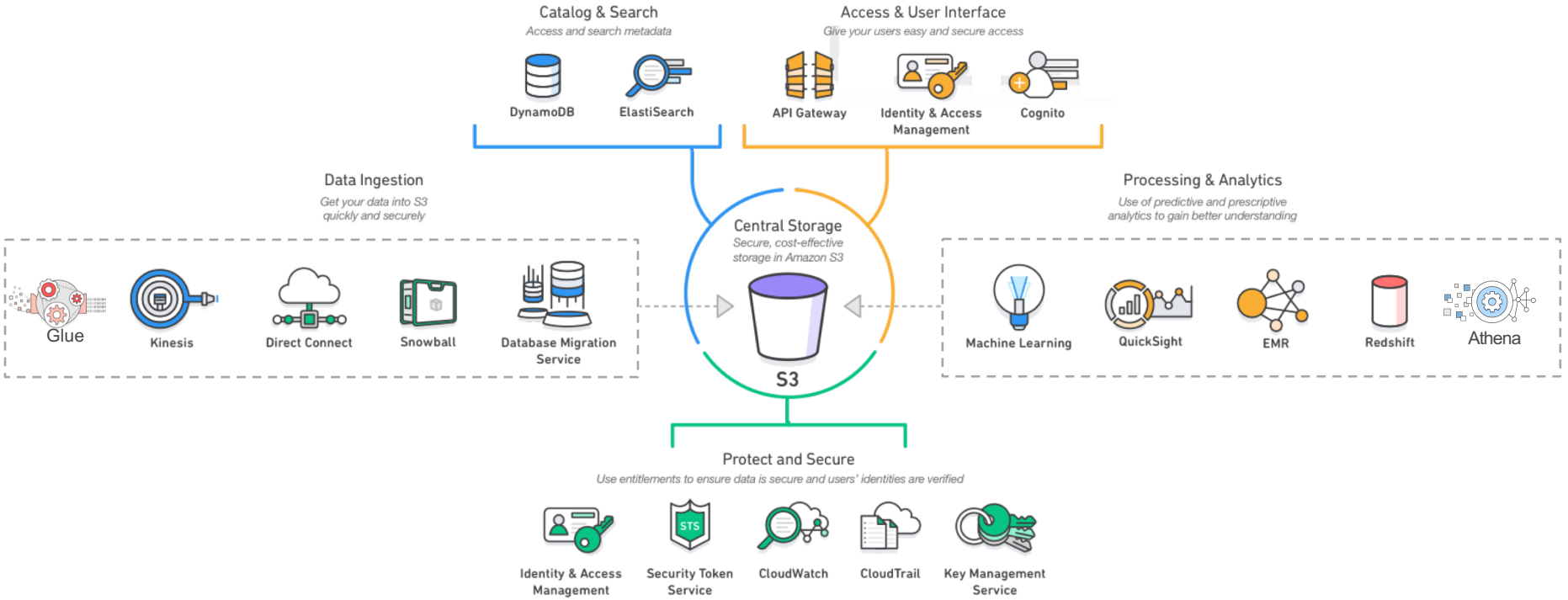
Dive in Anywhere



Flexible Access

Future Proof

Sample Reference Architecture: Data Lake



Get Your Hands Dirty Now?

Lab 1

https://d205nfn136cj6z.cloudfront.net/publicfiles/labguides/bigdata/glue-data-pipelines-2018/rudisur-aws-innovate-2018-glue-datapipelines-demo.htm

https://bit.ly/2FG5CxB

aws

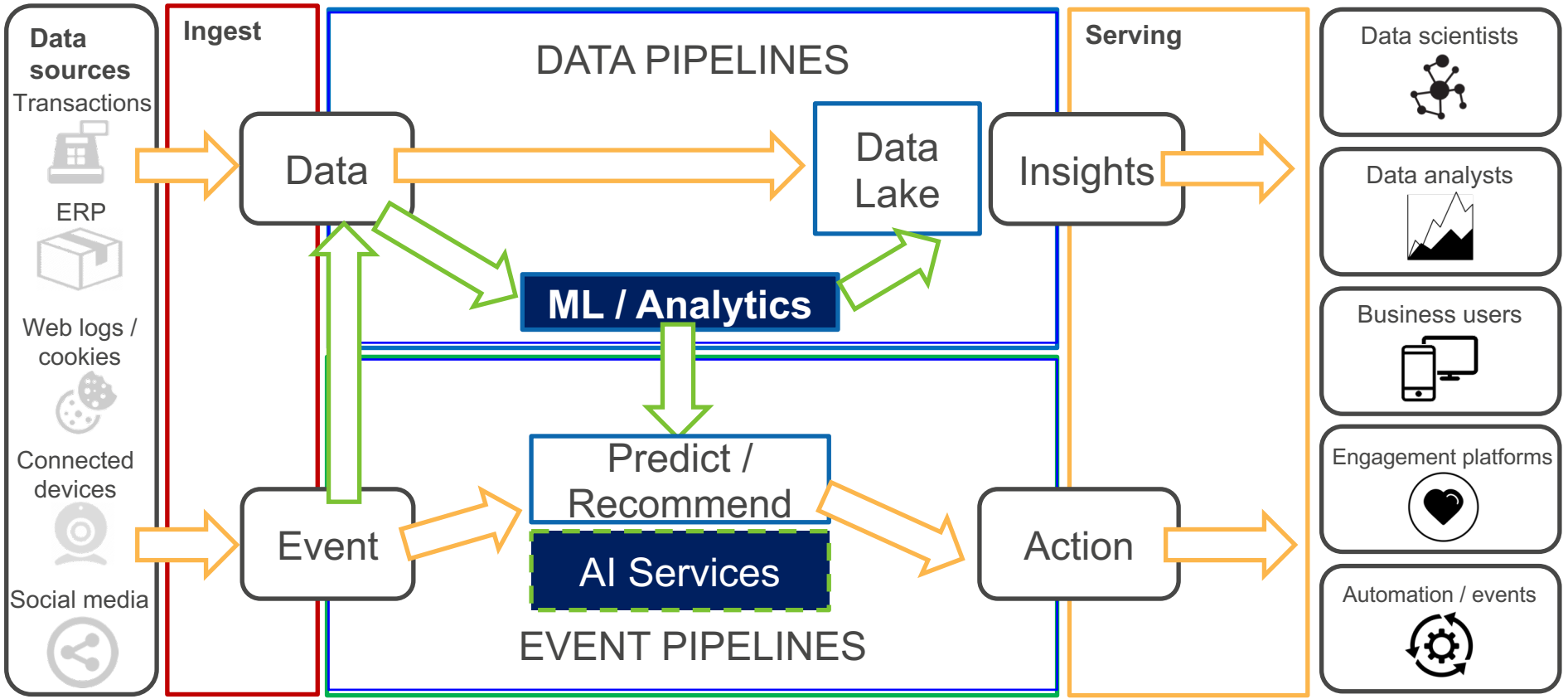
Modern Data Architecture

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

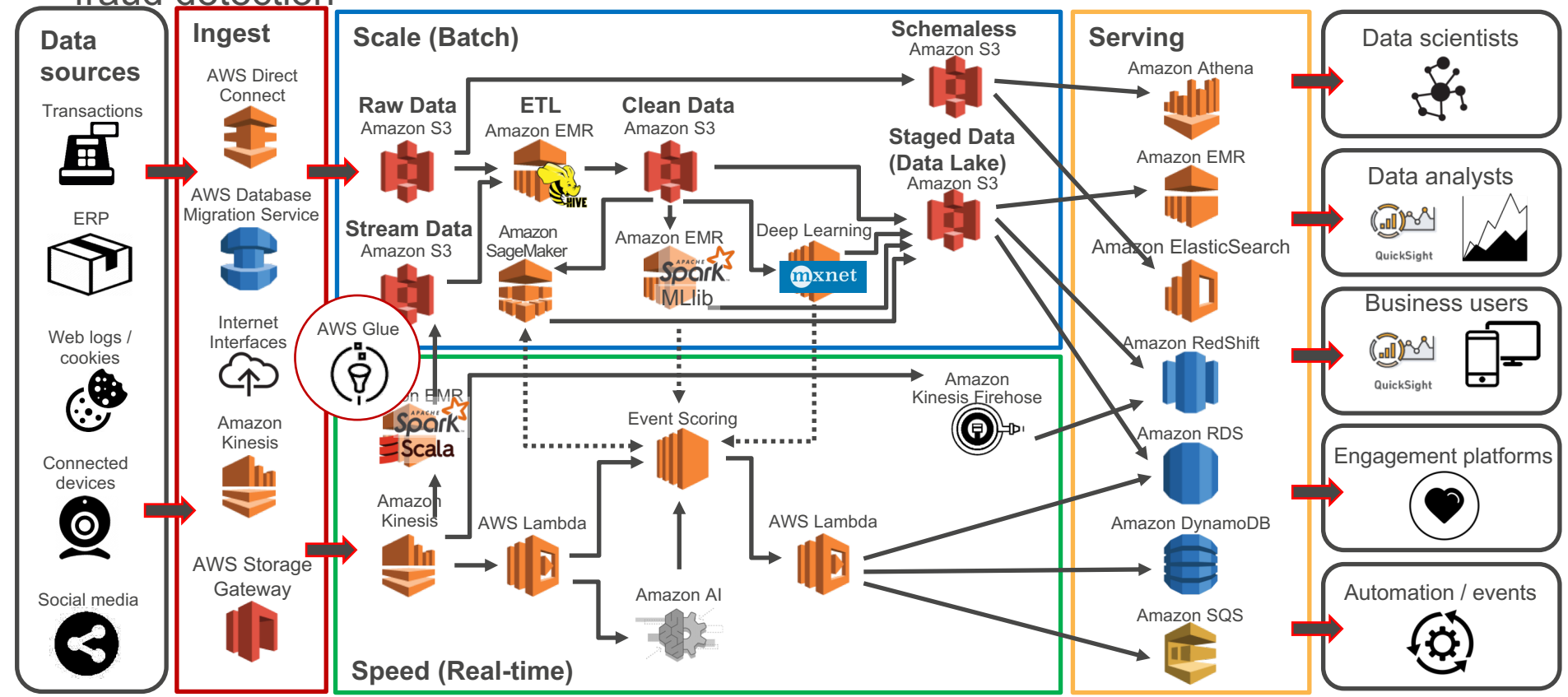
Modern data architecture

Insights to enhance business applications, new digital services



Real-time engagement

Interactive customer experience, event-driven automation, fraud detection



Get Your Hands Dirty Now?

Lab 2

https://dataprocessing.wildrydes.com/setup.html

Get Your Hands Dirty Now?

Lab 3

Go to Qwiklabs.com

Run lab

“Introduction to Amazon Redshift”

Getting Started: Tutorials & Blog

Try AWS with 10-Minute Tutorials

10-Minute Tutorials are simple "Hello, World!" technical documents to help you get hands-on with AWS.

10-Minute Tutorial
Launch a Linux VM
using Amazon EC2

10-Minute Tutorial
Store and Retrieve a File
with Amazon S3

10-Minute Tutorial
Launch a WordPress Website
with Amazon EC2 and AWS Marketplace

10-Minute Tutorial
Launch a Web Application
with AWS Elastic Beanstalk

10-Minute Tutorial
Register a Domain Name
using Amazon EC2

10-Minute Tutorial
Store Multiple Files
to Amazon S3 using the AWS CLI

10-Minute Tutorial
Update a Web Application
with AWS Elastic Beanstalk

10-Minute Tutorial
Create and Query a NoSQL Table
with Amazon Dynamo DB

Subscribe to the AWS Big Data Blog: http://blogs.aws.amazon.com/bigdata/

aws

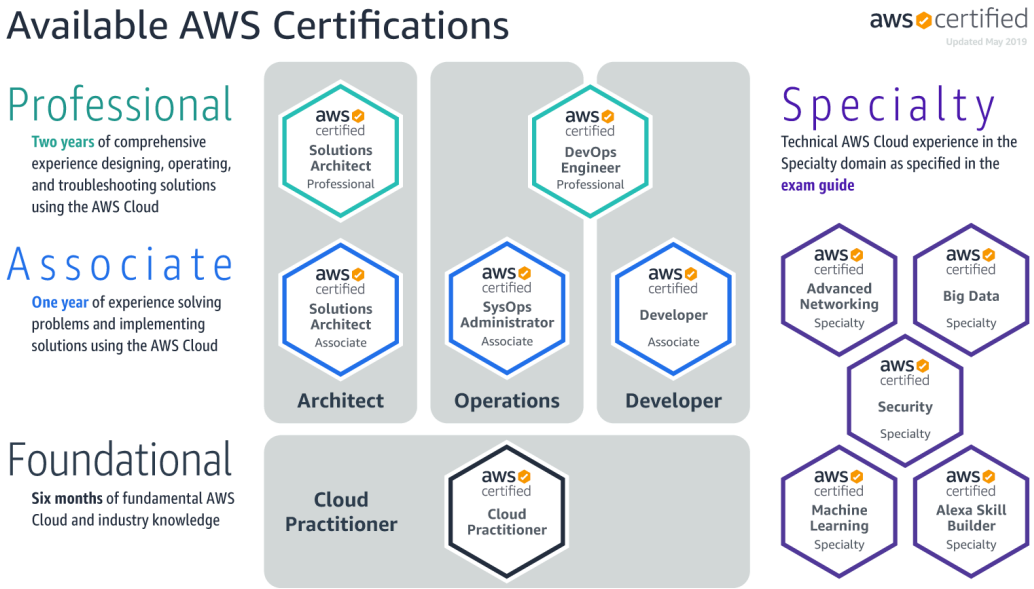
Summary

- Build sophisticated Big Data applications cost-effectively and support retrospective, real-time and predictive analysis
- You can build incrementally, scale automatically and add use cases as you go
- AWS delivers added benefits of security and auditing features to enable you to meet your stringent requirements
- Build hybrid applications that span across your datacenters and the AWS Cloud

aws

What's Next ?

1. Enroll to AWS Free Digital Training at https://aws.training
2. Get Train in Authorized Training Partner
3. Demonstrate you skill, proof it by getting AWS Certification



What's Next ?

Tell Your Friends to Attend Regular Workshop in AWS Office

A. Beginner Level:
    1) Launch Your First Workload on AWS

B. Intermediate Level
    1) Serverless on AWS
    2) Container on AWS
    3) Big Data and AI/ML on AWS

Samples of Analytic Architecture (Appendix)

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

A Sample Batch Analytics Pipeline

Upload data from multiple sources into S3

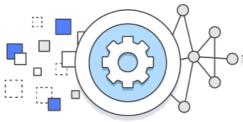
Use Amazon EMR to transform and cleanse the data (ETL)

Load formatted and cleansed data into S3

Redshift loads data in parallel optimizing it for fast analytics queries

Analyze and visualize data with Amazon Quicksight

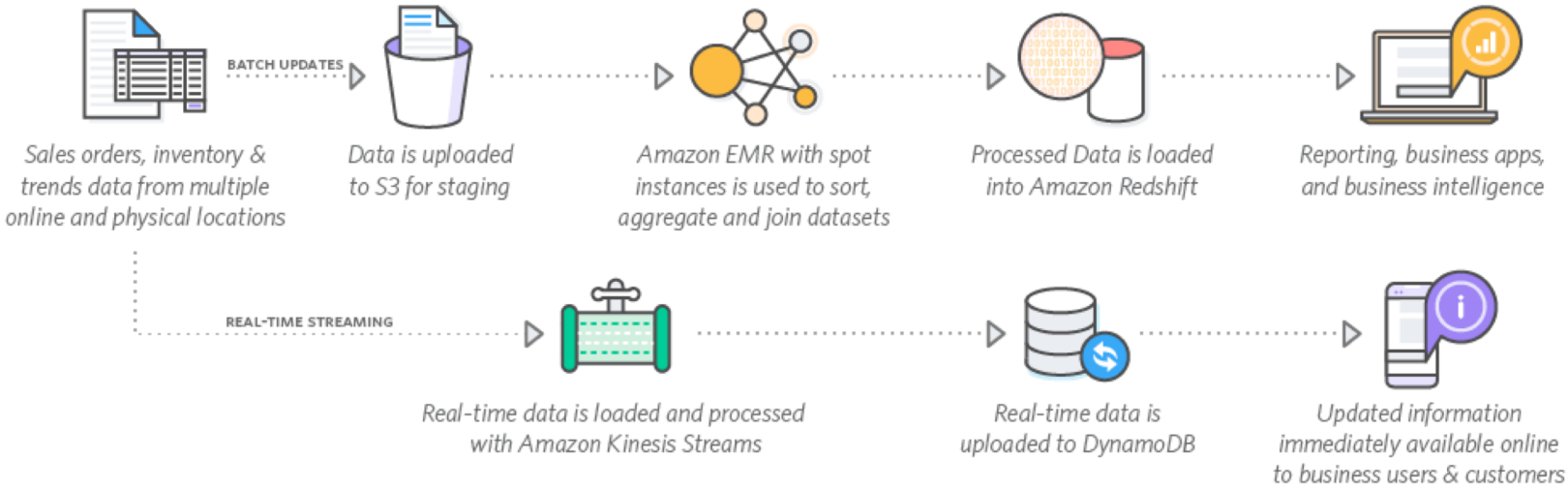
Ad-hoc access to data using Athena



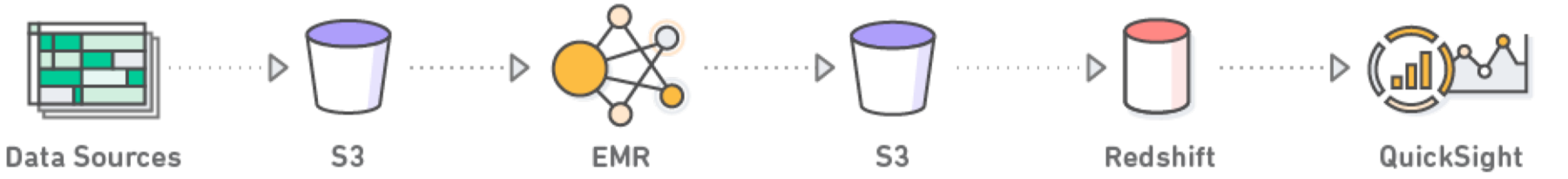
Athena can query aggregated datasets as well

aws

On-demand Big Data Analytics

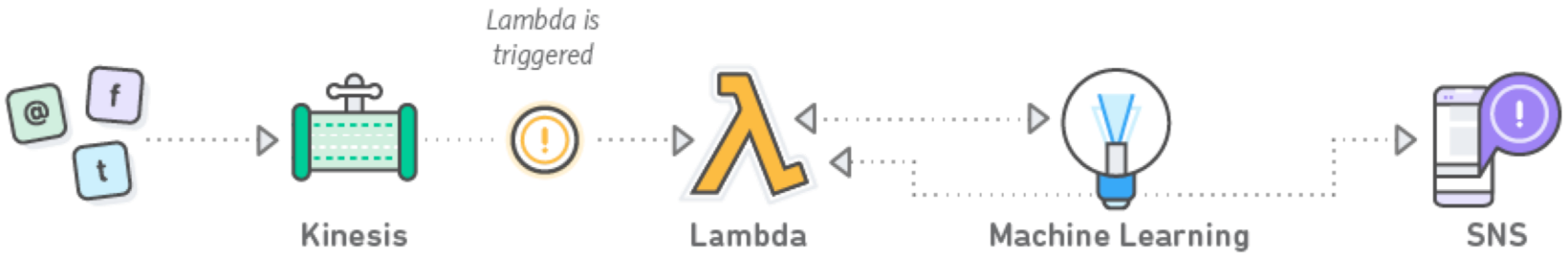


Data Warehousing



aws

Smart Applications | Machine Learning



Clickstream Analysis

Send clickstream data to Kinesis Streams

Kinesis Streams stores and exposes clickstream data for processing

Custom application built on Kinesis Client Library makes real-time content recommendations

Readers see personalized content suggestions

aws

Event-driven Extract, Transform, Load (ETL)